UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE E TECNOLOGIE

Corso di Laurea Triennale in Fisica

THE IMPACT OF HADRONIC OBSERVABLES ON THE GLUON DISTRIBUTION

Relatore:
Prof. Stefano Forte

Correlatore:
Dott. Juan Cruz-Martinez

Candidato:
Dario Chemoli
matricola n. 873024

Anno accademico 2021/2022

Abstract

The purpose of this thesis is to study the dependence of the gluon Parton Distribution Function on the proton-proton scattering data, as it could be the production of top and antitop quarks, jets, or the Z boson. We call these data processes hadronic observables.

Parton Distribution Functions (PDFs) cannot be computed from first principles: they have to be extracted from the data, through a careful comparison of theoretical predictions and experimental results. In order to determine the PDFs we use the NNPDF fitting methodology.

The NNPDF collaboration determines the structure of the proton using contemporary methods of artificial intelligence. NNPDF determines PDFs using as an unbiased modeling tool machine learning methods, which involve the application of neural networks, and use to construct a Monte Carlo representation of PDFs and their uncertainties: a probability distribution in a space of functions.

In Chapter 1 we summarize the main results of the Standard Model and QCD, and introduce the deep scattering processes and the scaling hypotesis. After we give this theoretical framework, we explain what is a PDF and how it appears in QCD computations, such as in the cross-sections formula. Then we briefly recall the recent history of PDFs determination.

In Chapter 2 we present the general strategy used in the NNPDF approach for the determination of a PDF set and in particular how the approach adopted in NNPDF4.0 fitting methodology (the newest methodology) represents an improvement made since the previous methodologies. Then we introduce the theoretical constraints which are imposed upon the PDF parameterization and how the neural network adopted in NNPDF4.0 fits the PDFs. Finally we talk about the global NNPDF4.0 fitting framework, and how the stochastic gradient descent method has a key role in determining the PDFs.

In Chapter 3 we present the results we got in the developed fit simulations. The basic idea of the work is the following: we produce a reference fit with experimental data taken from the complete NNPDF4.0 dataset; then we remove some data from the complete dataset which correspond to certain hadronic observables, comparing the new distribution fittings with the reference one, trying to understand how these observables have an impact on the gluon PDF.

Contents

Introduction

Particle physics is a branch of physics that studies the nature of matter and radiation: it investigates the irreducibly smallest objects that make up matter and the fundamental interactions necessary to explain their behaviour.

Particle physics is also called *high energy physics*, due to the high resolution needed to study elementary particles.

The concept of 'elementary' is used in the sense that such particles have no known structure, so they are pointlike. And the concept of pointlike depends on the resolution of the tools used to interact with the system which is studied: the resolution is Δr if two points of the system can just be resolved as separate when they are a distance Δr apart.

Assuming a probe is used in order to investigate the system, as it is used in scattering experiments, and assuming the probing beam itself consists of pointlike particles, the resolution is limited by the de Broglie wavelenght of these particles, which is $\lambda = \frac{h}{p}$, where p is the beam momentum and h is the well known Plank's constant. The beam of high momentum has short wavelenghts and then can have high resolution.

For example, the resolution of an optical microscope is given by

$$\Delta r \simeq \frac{\lambda}{\sin\theta} \quad , \tag{1}$$

where θ is the angular aperture of the light beam used to view the system. Then substituting the de Broglie relation, the resolution becomes

$$\Delta r \simeq \frac{\lambda}{\sin\theta} = \frac{h}{p\sin\theta} \simeq \frac{h}{q} \quad , \tag{2}$$

so that in a scattering experiment the resolution is inversely proportional to the momentum transferred q to the target [1].

So, the motivation for high resolution experiments in experimental particle physics is simply that many of the elementary particles are extremely massive, and the mass-energy mc^2 required to create them in scattering experiment is extremely large.

In this thesis we will use datasets from cross-section of high energy scattering experiments, in order to determine distribution functions of the partons contained in a proton, and then to gain information on the structure of this nucleon.

Chapter 1

Parton Distribution Functions

In this chapter we introduce the main subject of this thesis, the Parton Distribution Functions, and show how they are a fundamental tool to compute predictions at hadron colliders. First of all, however, we give some information about the hadronic cross-section, in the context of Quantum Chromodynamics.

1.1 Theoretical framework

1.1.1 Basics of Standard Model and QCD

All experimental data from high energy experiments can be accounted for by the *Standard Model* of particles and their interactions. According to this model, matter is built from a small number of fundamental spin $\frac{1}{2}$ particles, or *fermions*: six *quarks* and six *leptons*. For each of the various fundamental constituents, its symbol and the ratio of its electric charge q to the elementary charge q_e are given in Table 1.1. Moreover an antiparticle with opposite electrical charge can be associated to each fermion.

The leptons carry integer electric charge. The electron e is a familiar particle, while the other charged leptons are the muon μ and tauon τ: these are heavier versions of the electron. The neutral leptons are called *neutrinos*, indicated by the generic symbol ν. A different 'flavour' of neutrino is paired with each 'flavour' of the charged lepton.

The quarks carry fractional charges, of $\frac{2}{3}|q_e|$ or $-\frac{1}{3}|q_e|$. In Table 1.1, the quark masses increase from left to right, just as they do for the leptons. Moreover, just as for the leptons, the quarks are grouped into pairs differing by one unit of electrical charge. Sometimes quark are also called *partons*, because they constitute bigger particles, as the proton and the neutron.

While leptons exist as free particles, quarks do not. The fact that a free quark has never been observed is compatible with known properties of QCD, but cannot be proven from mathematical principles. No matter how energetically protons are collided together in accelerators: no quarks are seen to emerge in the debris. This phenomenon is known as *quark confinment*.

Particle	Flavour	$q/\|q_e\|$
leptons	e μ τ	-1
	ν_e ν_μ ν_τ	0
quarks	u c t	$+2/3$
	d s b	-1/3

Table 1.1: The fundamental fermions.

The Standard Model also describes the interaction between the fundamental fermions. The different interactions are described in terms of the exchange of characteristic *bosons*, which are integer spin particles, between the fermion constituents. These boson mediators are listed in Table 1.2.

Interaction	Mediator
strong	g
electromagnetic	γ
weak	$W^{\pm}$,Z^0

Table 1.2: The boson that mediate fundamental forces.

Excluding the gravitational field, which is not included in the Standard Model, there are three types of fundamental interactions, as follows [1].

Strong interactions are responsible for binding the quarks in the neutron and proton, and the neutrons and protons within nuclei. The interquark force is mediated by a massless particle, the *gluon*.

Electromagnetic interactions are responsible for virtually all the fenomena in extra-nuclear physics, in particular for the bound states of the electron with nuclei, and for the intermolecular force between liquids and solids. These interactions are mediated by the *photon* exchange.

Weak interactions determine the slow process of nuclear β-decay, involving the emission by a radioactive nucleus of an electron and a neutrino. The mediators of the weak interactions are the $W^{\pm}$ and Z^0 bosons, with masses of order 100 times the proton mass.

As remarked before, quarks do not exist as free particles. We must also say that the lightest bond states are the *baryon* and the *meson*, which are respectively a three quark state and a quark-antiquark pair. These strong interacting quark states are collectively referred to as *hadrons*. The fact that two and only two types of quark combinations occur is successfully accounted for in the theory of interquark forces, called Quantum Chromodynamics (QCD).

Quantum Chromodynamics is the theory of strong interactions between quarks and gluons. QCD is a quantum field theory known as a non-abelian gauge theory, with symmetry group SU(3). The QCD analog of electric charge is a property called colour. Quarks and gluons can be characterized by three types of colour charge: this colour charge is completely unrelated to the everyday meaning of colour. The term colour and the labels red, green, and blue became popular simply because of the loose analogy to the primary colors.

In this picture, all known particles have neutral colour charge: barions have three quarks of different colours (blue, red and green), while mesons have two quarks, one of one colour and the other of the relative anti-colour. The gluon carries out the interaction between coloured quarks, so it shows two coloured components, one colour and one anti-colour. In this way, there exist eight independent gluons, which can be combined to give different basis.

Protons and neutrons consist of the lightest u and d, three at a time: in particular a proton consist of uud, and a neutron consist of uud. The heavier quarks s, c, t, b also combine to form particles akin to, but much havier than, the proton and neutron, but these are unstable and decay rapidly (in typically 10^{-13} s) to u, d combinations, just as the heavy leptons decay to electrons. It is possible however to find also the haevier quarks inside the proton, because the interactions of the gluon.

Particle	Constituents	Mass [MeV/c^2]	Antiparticle	Constituents
Proton	uud	$938,28$	Antiproton	$\bar{u}\bar{u}\bar{d}$
Neutron	udd	$939,57$	Antineutron	$\bar{u}\bar{d}\bar{d}$
Gluon basis	$r\bar{b}$ $r\bar{g}$ $b\bar{g}$ $b\bar{r}$ $g\bar{r}$ $g\bar{g}$	$\frac{1}{\sqrt{2}}(r\bar{r}-b\bar{b})$	$\frac{1}{\sqrt{6}}(r\bar{r}+b\bar{b}-2g\bar{g})$	

Table 1.3: A table showing some properties of the proton, the neutron, their their respectively antiparticles, and a possible gluon basis, also known as *octet of states*.

1.1.2 Deep inelastic processes

Deep inelastic scattering is the name given to a process used to probe the insides of hadrons (particularly the baryons, such as protons and neutrons), using electrons, muons and neutrinos. It is an extension of Rutherford classic scattering to much higher energies of the probes and thus to much finer resolution of the components of the nuclei.

It is important that we introduce the concept of deep inelastic scattering, in order to introduce the parton model of the nucleon, which is the simplest model to factorize. Moreover we introduce the *Bjørken variable*, which we will see is a key quantity in the description of deep inelastic scattering processes with hadrons in the initial state.

The term *deep inelastic scattering* arises because the nucleon which is probed in the reaction nearly always disintegrates as a result. In particular this scattering process is named *inelastic* because this can be described as follows[1]:

$$e^- + p^+ \to e^- + X \quad , \quad \text{where} \quad M_X^2 > M_n^2 \quad .$$

In particular the inelastic scattering process is named *deep* if $M_X^2 \gg M_n^2$.

Deep inelastic scattering experiments divide into different classes, depending on the nature of the probe used, which indicates the force involved: in this case we talk about the electron-proton scattering, in which the leading process of the scattering is the one of a single photon exchange.

The main measurement of these scattering experiments is the cross-section, which is the effective target area of the nucleon, with the energy lost by the lepton during the collision and with the angle through which the incident lepton is scattered. The energy lost ν by the lepton is simply the difference between its incident and the final energy

$$\nu = E_i - E_f \, ,$$

while the angle through which the lepton is scattered is related to the square of the momentum transferred q^2 by the photon from the lepton to the nucleon, by the following relation:

$$q^2 = 2E_iE_f(1 - \cos\theta) \, .$$

These are two observables in deep inelastic scattering experiments, which connect the data from experiments with the theoretical picture of the proton interior.

Figure 1.1 shows the deep electron-proton inelastic scattering. In this diagram the incoming electron and proton have momentum respectively $\mathbf{k}$ and $\mathbf{p}$. The transferred momentum by the emitted photon is $q = |\mathbf{k} - \mathbf{k}'|$, and X is the final state of the proton scattered.

The deep inelastic scattering cross-section formula for the process in Figure 1.1 has to be characterised by some 'structure functions', which encode the structure of the proton. In particular, the perturbative methods of Quantum Electrodynamics (QED) show that the formula

[1] M_n and M_X are the nucleon mass and the final product mass

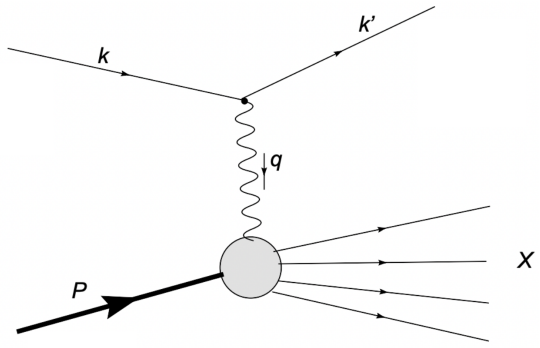


Figure 1.1: Feynman diagram for deep inelastic electron-proton scattering [2].

describing the differential cross-section for the electron-proton deep inelastic scattering with respect to the momentum transferred squared q^2 and the energy lost by the electron ν is the following [3]:

$$\frac{d^2\sigma}{dq\,d\nu} = \frac{4\pi\alpha^2}{q^4}\frac{E_f}{E_i M_p}\left[\frac{M_p}{\nu}F_2(q^2,\nu)\,\cos^2(\theta/2) + 2F_1(q^2,\nu)\,\sin^2(\theta/2)\right], \tag{1.1}$$

where M_p is the mass of the proton and α is the fine-structure constant.

However, in very high energy deep inelastic scattering processes, in which $q^2, \nu \to \infty$, the wavelenght of the photon is so small that the existence of the complete proton is really irrelevant to the reaction: the photon interacts with only a small part of it, and does so independently of the rest of it. This means that there is no justification for using the proton mass to determine the scale of the deep inelastic regime, so we can neglet the proton mass. Then it can be assumed that the structure functions depend only on a dimensionless ratio depending only by the quantities $|\mathbf{p}|$ and $|\mathbf{q}|$. Choosing such a ratio as

$$x = -\frac{q^2}{2\mathbf{p}\cdot\mathbf{q}} = \frac{Q^2}{2\mathbf{p}\cdot\mathbf{q}} \quad , \tag{1.2}$$

then the hypothesis which is made, known as *scaling hypotesis*, is that the structure functions can depend only on the dimensionless variable x, and not on either or both of the quantities p, q solved separately. So as $q^2, \nu \to \infty$ we have

$$F_{1,2}(q^2,\nu) \longrightarrow F_{1,2}(x) \quad . \tag{1.3}$$

In this framework the variable x, which is known as *Bjørken variable*, has a very significant interpretation. It turns out to be the fraction of the momentum of the proton carried by the parton which is struck by the photon. So the structure functions, which depend only on the parameter x, effectively measure the way in which the proton momentum is distributed among its constituents.

1.2 Factorization in QCD

In this section we present a basic property holding in QCD, that is the factorization, which will lead us to the definition of a PDF.

First of all we must say that QCD factorization leads to a factorization for the deep inelastic structure function (1.3)

$$F_i(x) = x\sum_j \int_x^1 \frac{1}{z}C_{ij}\Big(\frac{x}{z},\alpha_s(Q^2)\Big)\,f_j(z,Q^2)\,dz \tag{1.4}$$

where x is the standard Bjørken variable, C_{ij} is the structure function of a nucleon computed with an incoming parton, and $f_j(z, Q^2)$ is the distribution of the parton j in the only incoming nucleon.

Then factorized structure functions are used in QCD factorization in order to compute the cross-section for a generic hadroproduction process which depends on a single scale M_X^2. It can be in fact written in factorized form as [4]

$$\sigma_X(s, M_X) = \sum_{a,b} \iint_{x_{min}}^1 f_{\frac{a}{h_1}}(x_1, M_X)\; f_{\frac{b}{h_2}}(x_2, M_X)\, \sigma_{ab\to X}(x_1, x_2, s, M_X)\, dx_1\, dx_2 =$$
$$= \sum_{a,b} \sigma^0_{ab} \int_\tau^1 \frac{1}{x_1} f_{\frac{a}{h_1}}(x_1, M_X^2)\, dx_1 \int_{\frac{\tau}{x_1}}^1 \frac{1}{x_2} f_{\frac{b}{h_2}}(x_2, M_X^2)\, C_{ab}\Big(\frac{\tau}{x_1 x_2}, \alpha_S(M_X^2)\Big)\, dx_2 \qquad (1.5)$$

where s is the center-of-mass energy of the hadronic collision, $f_{\frac{a}{h_i}}(x_i, M_X^2)$ is the distribution of partons of type a in the i^{th} incoming hadron, $\sigma_{ab\to X}$ is the parton-level cross section for the production of the desired final state X, and the minimum value of x_i is

$$\tau = \frac{M_X^2}{s} \quad , \qquad (1.6)$$

where τ is called *scaling variable of the hadronic process*. The hard coefficient function $C_{ab}\big(z, \alpha_S(M_X^2)\big)$ is a function of the scale M_X^2 and the dimensionless ratio of this scale to the center of mass energy $\tilde{s}$ of the partonic subprocess

$$z = \frac{M_X^2}{\tilde{s}} = \frac{\tau}{x_1 x_2} \quad . \qquad (1.7)$$

In (1.5) a prefactor σ^0_{ab} has been extracted, so that at leading perturbative order (LPO) the coefficient function is either zero, for partons that do not couple to the given final state at leading order, or else just a Dirac delta:

$$\sigma_{ab\to X} = \sigma^0_{ab}\, C_{ab}\big(z, \alpha_S(M_X^2)\big) \quad , \quad C_{ab}\big(z, \alpha_S(M_X^2)\big) = c_{ab}\delta(1-z) + o(\alpha_S)$$

where the matrix C_{ab} depends on the specific process.

So, the importance of Parton Distribution Functions is that they encode the structure of strongly-interacting hadrons in the beam of high-energy collisions. We must underline that PDFs are not probability density functions, because they are not functions themselves, but distributions, and they are also not positive-definite. Finally we must observe that a single PDF depends on the mass-energy of the interacting parton in the collision and on the fraction x of the momentum of the parton inside the relative hadron, that is the Bjørken variable.

The factorized equation (1.5) express the hadronic cross section in terms of PDFs at the same scale at which the hadronic cross-section is evaluated, namely a reference value $M_X^2 = Q_0^2$. However, PDFs at different scales are related by perturbative evolution equations, namely the integro-differential equations

$$\frac{\partial}{\partial(\ln Q^2)} \begin{pmatrix} \Sigma(x, Q^2) \\ g(x, Q^2) \end{pmatrix} = \int_x^1 \begin{pmatrix} P^S_{qq}\Big(\frac{x}{y}, \alpha_S(Q^2)\Big) & 2n_f P^S_{qg}\Big(\frac{x}{y}, \alpha_S(Q^2)\Big) \\ P^S_{gq}\Big(\frac{x}{y}, \alpha_S(Q^2)\Big) & P^S_{gg}\Big(\frac{x}{y}, \alpha_S(Q^2)\Big) \end{pmatrix} \begin{pmatrix} \Sigma(x, Q^2) \\ g(x, Q^2) \end{pmatrix} \frac{dy}{y}\,, \quad (1.8)$$

$$\frac{\partial}{\partial(\ln Q^2)} q^{NS}_{ij}(x, Q^2) = \int_x^1 P^{NS}_{ij}\Big(\frac{x}{y}, \alpha_S(Q^2)\Big)\, q_{ij}(y, Q^2)\, \frac{dy}{y}\,, \qquad (1.9)$$

where $g(x, Q^2)$ is the gluon distribution, while $\Sigma(x, Q^2)$ denotes the singlet quark distribution defined as

$$\Sigma(x, Q^2) = \sum_{i=1}^{n_f} \left(q_i(x, Q^2) + \bar{q}_i(x, Q^2) \right) \quad , \tag{1.10}$$

where n_f is the number of quark flavours which we want to consider, while the nonsinglet quark distributions are defined as any linearly independent set of $2n_f - 1$ differences of quark and antiquark distributions

$$q_{ij}^{NS}(x, Q^2) = q_i(x, Q^2) - \bar{q}_j(x, Q^2) \quad , \tag{1.11}$$

and lastly the splitting functions P_{ab} are perturbative series in α_S, that start at order α_S at Leading Order (LO).

Finally we must say that perturbative evolution has some constrains due to conservation laws, one of which is the conservation of the total energy-momentum, which impose

$$\int_0^1 \left[\sum_{i=1}^{n_f} q_i(x, Q^2) + \bar{q}_i(x, Q^2) + g(x, Q^2) \right] x \, dx = 1 \quad , \tag{1.12}$$

while the other theoretical costraints are discussed in Section 2.2.

Therefore, combining the factorized expression in equations (1.5) with the solution of the integro-differential equations (1.8) and (1.9), it is possible to find a way in which the PDFs can be computed starting from the experimental values of the hadronic cross-section, from the fitted PDFs in a reference value Q_0^2.

1.2.1 The determination of PDFs

As we can see from equation (1.5), the PDFs depend on the fraction x of the momentum assumed by the interacting parton inside the relative hadron, and on the mass of the final product obtained by the scattering hadrons. The dependence of the single PDF on the parameter Q is determinable through the set of differential equations (1.8) and (1.9). The dependence of the PDF on the parameter x, however, is not determinable in the same way as the dependence on Q: in order to determine this dependence we should operate in the non-perturbative domain of QCD, but this would imply the knowledge of the wavefunctions of the single partons.

First of all then, we must consider a set of Parton Distribution Functions evaluated in a reference value Q_0^2 and only depending on the Bjørken variable

$$f_i = f_i(x, Q_0^2) \quad , \tag{1.13}$$

where $0 < x < 1$, and we need a methodology in order to fit this set of functions, starting from the NNPDF4.0 complete cross-section dataset.

We need to highlight that the index i in Equation (1.13) stands for the i^{th} PDF 'flavour': we have already seen that in Equation (1.5) a PDF is defined by the scattering hadron and the relative interacting parton. Here we will consider only incoming nucleons, so the incoming hadron can be considered to be a proton.

In principle we can identify 13 independent PDFs (12 for quarks and anti-quarks and 1 for the gluon). However, top quark and bottom quark are assumed to be perturbatively generated, so we are going to fit just the four lighter quark PDF flavours, the corresponding anti-quark flavours, and the gluon: the PDFs which we are going to fit are then denoted as $\{u, \bar{u}, d, \bar{d}, s, \bar{s}, c, g\}$. We also assume charm and anti-charm PDF to be the same, so in practice we fit 8 independent PDFs.

One of the first modalities in which we can approach the problem of fitting the PDFs is postulating a parameterization of the single PDF through the choice of free parameters, for example:

$$f_i(x, Q_0^2) = x^{\alpha_i}(1-x)^{\beta_i} \quad . \tag{1.14}$$

The parameterization of the single PDF shown in Equation (1.14) is clearly very simple: choosing this parameterization or some other more complicated parameterization can lead us to some advantages but also to some disadvantages: for example considering the parameterization (1.14) the only parameters we have to compute per PDF flavour are α_i and β_i, and in order to compute these free parameters we need a relatively small amount of data. However, in assuming such parameterization, the analysis we make is certainly subject to a bias. Moreover, in case the parameterization (1.14) doesn't lead to a correct result, even if we choose a more complicated parameterization (determined by a greater number of parameters), the analysis we make will always be subject to a bias.

So, in order not to introduce a bias, we use machine learning methods in which, starting with the experimental data, it is possible to extract quantities (in this case the PDFs) that depend on them.

In Figure 1.2 we present an example of PDFs fitted by the NNPDF4.0 machine learning fitting methodology at $Q = 3.2\,\text{GeV}$ and $Q = 10^2\,\text{GeV}$, in which we consider Next to Next Leading Order accurancy (NNLO).

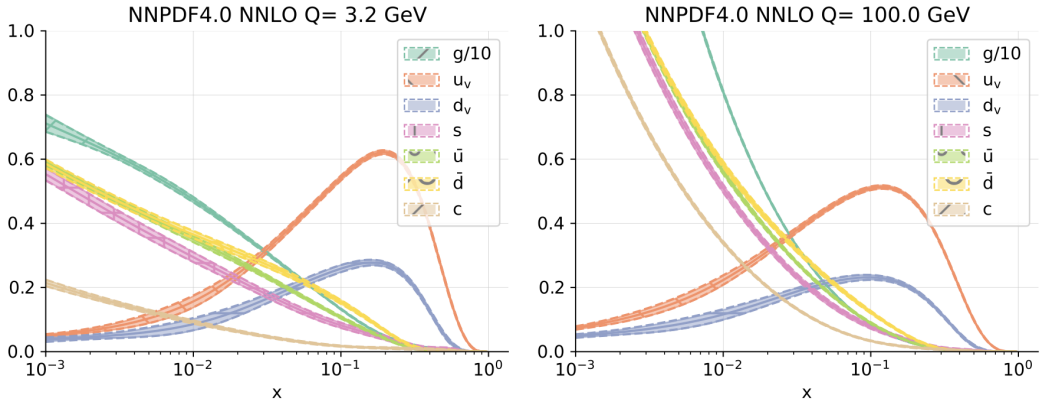


Figure 1.2: The NNPDF4.0 NNLO accurancy PDFs at $Q = 3.2\,\text{GeV}$ (left) and $Q = 10^2\,\text{GeV}$ (right) [5].

Chapter 2

The fitting framework

In this chapter we show the general strategy used in the NNPDF approach in order to fit the Parton Distribution Functions starting from a cross section dataset, and especially the NNPDF4.0 global dataset and machine learning techniques.

First of all we must introduce the improvements of NNPDF fitting methodology, comparing the NNPDF3.1 and NNPDF4.0 methods, and explaining how the last methodology leads to a better set of PDFs.

The NNPDF3.1 fitting methodology was the first to extensively include LHC data, and was able to reach 5% precision in PDF uncertainties. The machine learning technique used by NNPDF3.1 in order to fit the PDFs is based on using *genetic algorithms* [6]. The NNPDF4.0 fitting methodology represents a major step forward in many significant aspects, such as the systematic inclusion of an extensive set of LHC scattering processes at $7, 8$ TeV data and, for the first time, the inclusion of LHC data at 13 TeV and of several new processes not considered before for PDF determinations. Moreover the NNPDF4.0 machine learning technique is considerably faster and leads to more precise PDF fittings, due to the use of *stochastic gradient descent methods*, provided by `TensorFlow` library [7]. In fact NNPDF4.0 is the first PDF determination methodology based on a fitting that is selected automatically rather than through manual iterations and human experience.

2.1 Overview on NNPDF4.0 dataset

The global NNPDF4.0 dataset builds upon NNPDF3.1, by adding various new datasets to it, which is a variety of new LHC measurements for processes already present in NNPDF3.1 on the one hand, and data corresponding to new processes on the other. New datasets for existing LHC processes are added for electroweak boson production, both inclusive and in association with charm, single-inclusive jet production, and top pair production. The new processes are gauge boson with jets, single top production, inclusive isolated photon production, and dijet production [5].

Now we present in detail a list of the new datasets considered in NNPDF4.0 regarding LHC cross section data. The LHC cross section data have in fact a particular importance in this thesis.

- **Inclusive collider electroweak gauge boson production**: we include the ATLAS measurements of the W and Z differential cross-section at 7 TeV in the central and forward rapidity regions. These data were already included in NNPDF3.1, but only the subset corresponding to the central region [8];

- **Gauge boson production with additional jets**: on top of inclusive gauge boson production, we consider more exclusive measurements in which a W boson is produced in association with n jets of light quarks, or with a single jet of charm quarks. Specifically, we include the ATLAS data for W production with $n \geq 1$ at 8 TeV [9];
- **Top pair production**: we consider several new datasets for top pair production at the LHC. At 8 TeV, we include the ATLAS normalized differential cross-section and the CMS normalized double differential cross-section, both of which are measured in the dilepton channel [10],[11];
- **Single-inclusive and dijet production**: For single-inclusive jet production, we include the ATLAS and CMS measurements at 8 TeV [12];
- **Inclusive isolated-photon production**: Isolated photon production was not included in previous NNPDF releases and is included in NNPDF4.0 for the first time. We specifically consider the ATLAS measurements at $8, 13$ TeV [13],[14];
- **Single top production**: Another process included for the first time in an NNPDF release is t-channel single top production. We consider ATLAS and CMS measurements at $7, 8, 13$ TeV [15],[16],[17],[18],[19],[20].

2.2 Overview on fitting methodology

As we said before, NNPDF4.0 is the first fitting methodology fully selected through a machine learning algorithm. This means that the use of a Monte Carlo representation of PDF uncertainties and correlations, the use of neural networks as interpolating functions, and the choice of neural network architecture are now selected through an automated hyperoptimization procedure.

2.2.1 The general NNPDF approach

The PDFs determination can be seen as a *pattern recognition problem*, which is the determination of a set of functions starting from some data, and initially being unaware of the functional form of these functions. In this specific case, the pattern recognition problem is not one of the simplest, since the available data are not values of PDFs, but partonic cross-section values, which depend in a non-linear way on the PDFs themselves.

Another peculiarity of this pattern recognition problem is that the objects needed to be determined are not measurable quantities, but probability distributions. So through a statistical analysis, from the cross-section data, we can calculate probability density functions of the PDFs, which are probability density functions of probability density functions, also known as *probability functionals*.

The fundamental idea behind the general NNPDF approach, shown in Figure 2.1, is the following. From the hadronic cross-sections datasets, Monte Carlo methods can be used in order to build a multigaussian in the space of data using mean values and variances. By building the multigaussian it is possible to generate as many data replicas $\mathcal{D}_i$ as we want. Then every data replica is used to fit the most probable PDF flavours by using a neural network. Finally by using the fitted PDFs, it is possible to make predictions $\mathcal{O}_i$ on the data.

Iterating the process, and comparing the data $\mathcal{D}_i$ with their predictions $\mathcal{O}_i$ it is possible to build the *loss function*

$$\chi^2 = \frac{1}{N} \sum_{i,j=1}^{N} (\mathcal{D}_i - \mathcal{O}_i) M_{ij}^{-1} (\mathcal{D}_j - \mathcal{O}_j) \quad , \tag{2.1}$$

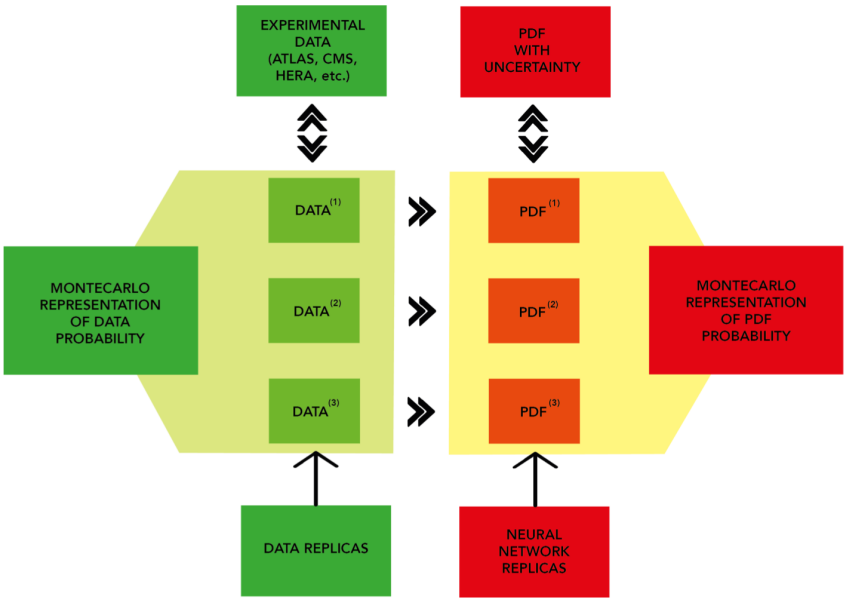


Figure 2.1: General strategy used in the NNPDF approach [6].

where the value M_{ij} is the component of the covariant matrix between data points i and j, and N is the total number of data. Finding the best PDFs is then possible by the minimization of the loss function.

The process we have just described can be used for every set of data replica, for which we have a set of PDFs fitted. Then the Monte Carlo methods are once again used in order to build the PDF distributions, using each set of PDF fitted by the datasets.

So given a dataset, for every data the neural network fits the most probable PDF $f_k(x, Q_0^2)$ evaluated in the reference value Q_0^2. An example of neural network used in NNPDF approach in determining a single PDF flavour is shown in Figure 2.2: this type of neural network is also known as *2-5-3-1 structure neural network*. The input of this neural network is the couple $(x, \ln x)$, and this is due to the two different regimes of PDFs: they show a linear regime for $0.03 \leq x \leq 0.5$, and a logarithmic regime in $10^{-4} \leq x < 0.03$. The output of the neural network is then the product $xf_k(x, Q_0^2)$, which covers a smaller range of values than $f_k(x, Q_0^2)$, so it is easier to show graphically.

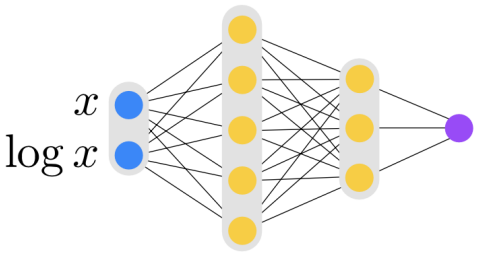


Figure 2.2: Example of the neural network used in NNPDF approach [6].

2.2.2 PDF parameterization and sum rules

We now turn to the general structure of the PDF parameterization, and the theory constraints that are imposed upon it. In Section 1.2 we talked about the conservation of the total energy-momentum, summarised in equation (1.12). In this section specifically we talk about sum rules, and PDF positivity and integrability.

The PDF fitting methodology requires a choice of basis, namely a set of linearly independent PDF flavour combinations, which are parameterized at the reference value Q_0^2. In the NNPDF approach, this corresponds to choosing the PDF combinations whose value is the output of a neural network. Results should in principle be independent on this specific choice of basis.

A possible set of linearly independent flavour combinations is defined *flavour basis*, and it contains the single parton PDF set:

$$\tilde{f}_k = \{u, \bar{u}, d, \bar{d}, s, \bar{s}, c, g\} \quad .$$

However, the default choice of PDF basis in NNPDF, called *evolution basis*, is the following:

$$f_k = \{V, V_3, V_8, T_3, T_8, T_{15}, \Sigma, g\} \quad ,$$

in which the basis PDFs are chosen as the singlet quark Σ and gluon g distribution, the valence V_i and nonsinglet sea T_i combinations that are eigenstates of QCD evolution, namely

$$\begin{aligned}
\Sigma &= u + \bar{u} + d + \bar{d} + s + \bar{s} + 2c \ , \\
T_3 &= (u + \bar{u}) - (d + \bar{d}) \ , \\
T_8 &= (u + \bar{u} + d + \bar{d}) - 2(s + \bar{s}) \ , \\
T_{15} &= (u + \bar{u} + d + \bar{d} + s + \bar{s}) - 3(c + \bar{c}) \ , \\
V &= (u - \bar{u}) + (d - \bar{d}) + (s - \bar{s}) \ , \\
V_3 &= (u - \bar{u}) - (d - \bar{d}) \ , \\
V_8 &= (u - \bar{u} + d - \bar{d}) - 2(s - \bar{s}) \ .
\end{aligned} \tag{2.2}$$

The evolution and flavour bases each have advantages and disadvantages. For instance, if we choose a factorization scheme in which PDFs are non-negative [21], positivity is easier to implement in the flavour basis. On the other hand, the integrability of the valence distributions V, V_3, V_8, as required by the valence sum rules, is simpler in the evolution basis. In this thesis, we take the evolution basis as our standard choice, as it is chosen in NNPDF4.0 methodology.

Once we have talked about the need for a choice of basis, we must say that the relationship between the neural network output and the PDFs is the following

$$x\, f_k(x, Q_0^2; \boldsymbol{\theta}) = A_k\, x^{1-\alpha_k} (1-x)^{\beta_k} \mathcal{N}_k(x; \boldsymbol{\theta}) \quad , \quad k = 1, \ldots, 8 \quad , \tag{2.3}$$

where k runs over the elements of the PDF flavour basis, $\mathcal{N}_k(x)$ is the neural network output, and $\boldsymbol{\theta}$ indicates the full set of neural network parameters. The constants α_k, β_k are parameters which are evaluated in every PDF fitting process: in order to make sure that the neural network does not bias the result, the parameters are varied in a range that is determined iteratively in a self-consistent manner [22]. The normalization constants A_k are constrained by the sum rules.

The sum rules determine the theoretical constraints on the PDFs. Irrespectively of the choice of the fitting basis, PDFs should satisfy both the momentum sum rule (1.12), and the

three valence sum rules:

$$\begin{aligned}\int_0^1 \left[u(x,Q^2)-\bar{u}(x,Q^2)\right]dx &= 2 \quad ,\\ \int_0^1 \left[d(x,Q^2)-\bar{d}(x,Q^2)\right]dx &= 1 \quad ,\\ \int_0^1 \left[s(x,Q^2)-\bar{s}(x,Q^2)\right]dx &= 0 \quad ,\end{aligned} \tag{2.4}$$

which express the conservation of baryon number.

The sums (1.12) and (2.4) must be valid for all values of Q^2. Provided that these sum rules are imposed at the initial parameterization scale Q_0^2, perturbative QCD ensures that they will hold for any other value.

Equations (1.12) and (2.4) are the flavour basis sum rules. When transformed to the evolution basis, the valence sum rules read

$$\int_0^1 V(x,Q_0^2)\,dx = \int_0^1 V_8(x,Q_0^2)\,dx = 3 \quad , \quad \int_0^1 V_3(x,Q_0^2)\,dx = 1 \quad . \tag{2.5}$$

while (1.12) remains the same.

Equations (1.12) and (2.5) fix four of the normalization constants A_k which appeared in (2.3), namely $A_g, A_V, A_{V_3}, A_{V_8}$, using the evolution basis.

In Figure 2.2 we have shown an example of the neural network used in NNPDF approach. After defining the flavour basis and the evolution basis we can now show in Figure 2.3 the true neural network architecture adopted for NNPDF4.0. In this case we remark that a single network is used, and its eight output values are the PDFs in the evolution basis (red box) or in the flavour basis (blue box). The main differences between the neural networks in Figure 2.2 and in Figure 2.3 are the number of hidden layers and the number of outputs. Both neural networks in fact have the same number of inputs, which are determined by the couple $(x, \ln x)$, but the one adopted in NNPDF4.0 has an additional hidden layer (in Figure 2.3 the hidden layers are labeled with $n^{(2)}$ and $n^{(3)}$): in statistical learning the *hidden layers* are simply layers of mathematical functions each designed to produce an output specific to an intended result. Moreover the neural network in Figure 2.2 has a single output, which is the single fitted PDF, whereas the NNPDF4.0 neural network has eight outputs, which correspond to the results shown in Figure 2.3.

2.2.3 Positivity and integrability

We now must talk about other two theoretical constrains: positivity and integrability of PDFs. These constraints are extremely relevant in the analysis of the fitting framework.

The hadronic cross sections are a non-negative quantities, because they are probability distributions. On the other hand we have already said that the PDFs are not probability distributions, thus they may be negative. Whether they are positive or negative, it depends on the factorization scheme. It is possible to show [21] that PDFs for individual quark flavours and the gluon in the $\overline{\text{MS}}$ factorization scheme are non-negative. So we now also impose this positivity condition along with the constraint of positivity of physical cross-sections discussed above.

PDF positivity is implemented by means of Lagrange multipliers. Specifically, for each PDF flavour, one adds a contribution to the total cost function used for the neural network training

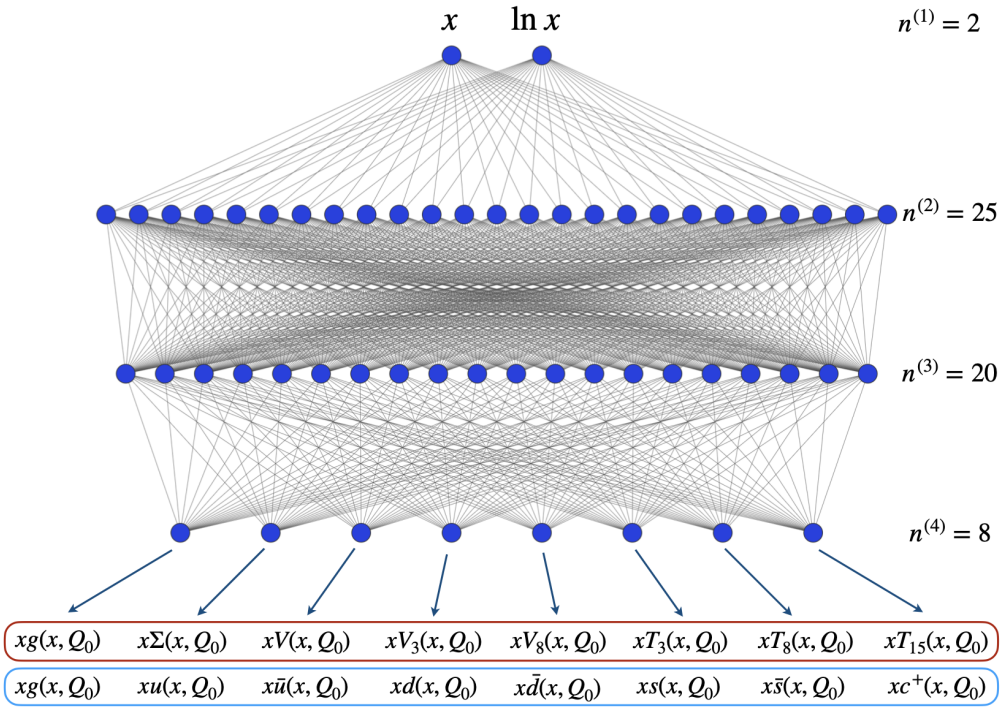


Figure 2.3: The neural network architecture adopted in NNPDF4.0 [5].

given by

$$\chi^2 \longrightarrow \chi^2 + \sum_{k=1}^{8} \Lambda_k \sum_{i=1}^{20} \Phi_\gamma\big(-f_k(x_i, Q')\big) \quad , \tag{2.6}$$

where $Q' = 5\,\text{TeV}$ and the x_i values are given by 10 points logarithmically spaced in $(5 \cdot 10^{-7}, 10^{-1})$ and 10 points linearly spaced in $(0.1, 0.9)$. The Φ_γ function is given by

$$\Phi_\gamma(t) = \begin{cases} t & \text{if} \quad t > 0 \\ \gamma(e^t - 1) & \text{if} \quad t < 0 \end{cases} \quad ,$$

with the parameter $\gamma = 10^{-7}$.

The equation (2.6) shows how the cost function bounded with a negative PDF receives a contribution which is proportional both to the corresponding Lagrange multipliers Λ_k and to the absolute magnitude of the PDF itself. This contribution will affect PDFs that assume negative values, which will thus not be considered as best fits.

In addition to the positivity requirement, small-x behavior of the PDFs is constrained by integrability requirements.

First of all, the gluon and singlet PDFs must satisfy the momentum sum rule (1.12), which implies that

$$\lim_{x\to 0} x^2 f_k(x,Q) = 0 \ , \ \ \forall Q \ , \ \ f_k = g, \Sigma \ \ , \tag{2.7}$$

while the valence sum rules (2.4) constrain the small-x behavior of the valence distributions

$$\lim_{x\to 0} x f_k(x,Q) = 0 \ , \ \ \forall Q \ , \ \ f_k = V, V_3, V_8 \ \ , \tag{2.8}$$

in the evolution basis.

Furthermore, the standard Regge theory suggests that [23]

$$\lim_{x\to 0} x f_k(x,Q) = 0 \ , \ \ \forall Q \ , \ \ f_k = T_3, T_8 \ \ . \tag{2.9}$$

In fitting the PDFs, the framework has to verify the conditions (2.7),(2.8), and (2.9) in order to ensure the integrability of the PDFs.

2.3 The state of the art

Finally, we talk about the global NNPDF4.0 fitting framework. This fitting framework can be divided into three main steps.

1. **Initialization**: inputs are given in order to initialize the neural network;
2. **Fitting and evaluation**: the neural network fits the PDFs from the data replica sets, and the algorithm evaluates the cost function and minimizes it;
3. **Post-fit selection**: the APFEL program evaluates the best PDFs at different values of Q^2 through the perturbation theory of QCD, then the PDFs are selected verifying their positivity and integrability. Finally, the output is made in the LHAPDF format in order to be viewed.

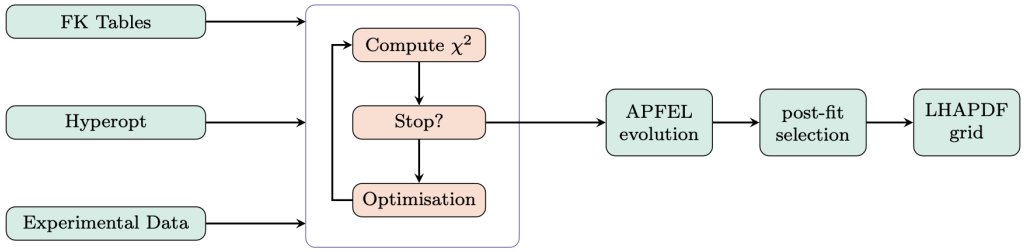


Figure 2.4: Diagrammatic representation of the NNPDF4.0 fitting framework [5].

2.3.1 Initialization

First of all, we have to give something to the machine in order to have something back (in this case the PDF fittings). It is obvious that the first thing to give the machine is the experimental data, which are organized in datasets.

The next step is the hyperoptimization procedure, or neural network training, which requires as input a number of methodological choices, such as the neural network architecture and the training rate. We can view these choices as the set of hyperparameters, which are denoted as $\boldsymbol{\theta}$ in (2.3). This set of hyperparameters defines a specific fitting strategy. While in many methodologies (including previous NNPDF determinations) these hyperparameters are determined by trial and error, in NNPDF4.0 an automated algorithmic procedure is implemented in order to scan the space of hyperparameters and determine the optimal configuration according to a figure of merit. In this work, the implementation of the hyperparameter scan is based on the `hyperopt` library [24].

Finally, in order to compute the loss function (2.1), the algorithm has to compute the hadronic cross section starting from the experimental data. And to do so, the algorithm also

has to compute the convolution integral (1.5) once the PDFs are fitted. Therefore, the PDFs are convoluted with partonic scattering cross-sections (including perturbative QCD evolution): the convolution integrals are substituted with convolution products which are encoded in pre-computed grids called FK-tables.

2.3.2 Fitting and evaluation

Once we give the input, the machine is ready to begin the process of fitting.

We have already said that the best way to fit the PDFs is using machine learning methods. In particular the problem that NNPDF4.0 fitting framework tries to solve is often known as *unsupervised problems* in machine learning: this kind of problems is based on having some dataset for which one wants to extract information or quantities that depend on the experimental data.

Previous NNPDF determinations used stochastic algorithms for the training of neural networks, and, in particular, in NNPDF3.1 nodal genetic algorithms were used. Stochastic minimization algorithms are less prone to end up trapped in local minima, but are generally less efficient than deterministic minimization techniques. In the approach adopted here in NNPDF4.0, the algorithms that we consider are *Stochastic Gradient Descent* algorithms implemented in the `Tensorflow` package [7].

Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. The idea is to take repeated steps in the opposite direction of the gradient of the function at the current point, because this is the direction of steepest descent. In this case the algorithm chooses a set of hyperparameters $\boldsymbol{\theta}_n$ in order to fit the PDF as in (2.3) for given data of a dataset. Then, as the algorithm iterates the fitting and calculate the loss function χ^2, it has to choose other hyperparameters in order to minimize the loss function. Considering that this function depends on the PDF fitted, and so it also depends on the hyperparameters, the new hyperparametrization chosen by the algorithm is such that

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \gamma_n \nabla_{\boldsymbol{\theta}} \chi^2(\boldsymbol{\theta}_n) \tag{2.10}$$

where the factor γ_n is known as *step size*, and it is possible to show that it is given by

$$\gamma_n = \frac{|(\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n+1}) \cdot (\nabla\chi^2(\boldsymbol{\theta}_n) - \nabla\chi^2(\boldsymbol{\theta}_{n+1}))|}{||\nabla\chi^2(\boldsymbol{\theta}_n) - \nabla\chi^2(\boldsymbol{\theta}_{n+1})||^2} \quad . \tag{2.11}$$

In doing so, for every dataset we find the better fitting that minimizes the loss function.

The use of gradient descent algorithms ensures greater efficiency, while the use of hyperoptimization guarantees the best methodology without underfitting or overfitting the data replicas.

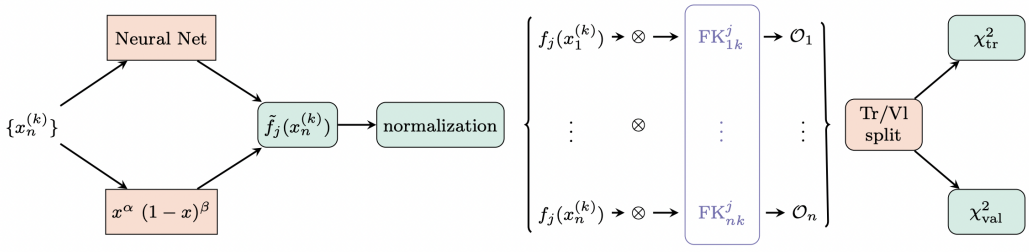


Figure 2.5: Diagrammatic representation of the calculation of the loss function in the NNPDF4.0 fitting framework [5].

Figure 2.5 illustrates the structure of the algorithm that evaluates the loss function (2.1) in terms of the PDFs fitted from the data replica.

In gradient descent method used in NNPDF4.0, a set of experimental data replica $\{x_n^{(k)}\}$ is generated[1] through Monte Carlo methods, to which a set of momentum fractions is associated. Then the code first fits the functions $\mathcal{N}_j(x;\boldsymbol{\theta})$ seen in equation (2.3) through the neural network and evaluates the preprocessing factors to construct un-normalized PDFs, which are then normalized determining the factor A_k, again from (2.3), in order to fit the every PDF flavour

$$f_{jn}^{(k)} \equiv f_j\Big(x_n^{(k)}, Q_0\Big) \quad , \tag{2.12}$$

where j, k, n in (2.12) label the j^{th} PDF flavour, the k^{th} data replica, and the n^{th} data from the data from the data replica. Then in order to evaluate the loss function χ^2, the algorithm associates a value on the fast Kernel tables to the $\{x_n^{(k)}\}$ data. This value is denoted as FK_n. Once the PDFs have been fitted by the neural network, the *4-rank luminosity tensor* is then defined as

$$\mathcal{L}_{i\alpha j\beta} = f_{i\alpha}\, f_{j\beta} \quad , \tag{2.13}$$

where (i, j) in (2.13) label the PDF flavour, while (α, β) label the preprocessed parameters.

Finally, the algorithm can determine a prediction of the x_n data $\mathcal{O}_n$ contracting the luminosity tensor with the n^{th} convolution product of the Fast Kernel tables

$$x_n \longrightarrow \mathcal{O}_n = \mathrm{FK}_n^{\ i\alpha j\beta} \mathcal{L}_{i\alpha j\beta} \quad , \tag{2.14}$$

and once the algorithms computes the set of reals $\{\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_n\}$, it is possible to determine the loss function χ^2. Then the algorithm iterates the procedure choosing a new set of parameters (2.10) in order to minimize the loss function.

Finally, in Figure 2.5 we can see that the loss function is subject to a split. The final product of the algorithm used in NNPDF4.0 are then the two loss functions χ^2_{tr} and χ^2_{val}. This is caused by the fact that, in fitting the PDFs, we may occur in the problem of fitting also the statistical noise.

Because of this noise, we cannot let the algorithm make too many iterations, because we may occur in the overfitting, that is the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably.

In order to remove the noise from fitting and in order to avoid the algorithm to overfit, it is introduced a *stopping criterion method*, and precisely the method used in this work is called *cross-validation method*: this method consists in splitting the experimental data in a certain dataset in two arbitrary sets, called *training set* and *validation set*, then compute the loss function χ^2 for every set. The ratio of these two sets is choosed a priori[2], but the data which are sent to these two sets are chosen randomly.

So, instead of having a single loss function, we have the two different loss functions

$$\chi^2_{tr} = \frac{1}{n} \sum_{i,j=1}^{n} (\mathcal{D}_i - \mathcal{O}_i) M_{ij}^{-1} (\mathcal{D}_j - \mathcal{O}_j) \quad ,$$

$$\chi^2_{val} = \frac{1}{m} \sum_{i,j=1}^{m} (\mathcal{D}_i - \mathcal{O}_i) M_{ij}^{-1} (\mathcal{D}_j - \mathcal{O}_j) \quad .$$

where n and m are the number of data contained in the training set and in the validation set respectively.

[1] The expression $\{x_n^{(k)}\}$ represents the n^{th} data from the k^{th} set of data replicas.

[2] In NNPDF4.0 usually the 75% of the data extracted from a dataset are sent in the training set, and the the 25% of the data are sent in the validation set.

The aim of the cross-validation method is that of minimize the loss function $\chi^2 = \chi^2_{tr}$, while the value of χ^2_{val} is monitored: the gradient descent method acts on the training set in order to minimize the function χ^2_{tr}, modifying the neural network weights as in equation (2.10). Simultaneously also the function χ^2_{val} changes because of the changes of the neural network weights. However, since the gradient descent method acts only on the training set, χ^2_{tr} might not always improve: so if this quantity stops improving, the process is arrested in order to avoid the overfitting. Since the noise has no correlation between the training set and the validation set, the cross-validation method ensures that the gradient descent method isn't fitting over the noise.

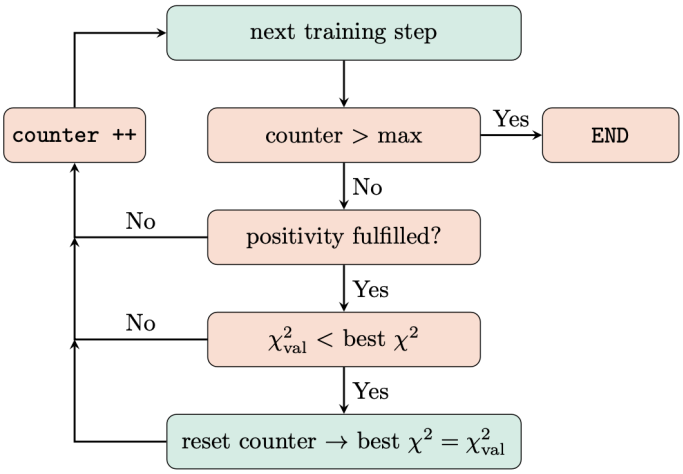


Figure 2.6: Flowchart describing the algorithm used in NNPDF4.0 to determine the optimal length of the iteration based on the cross-validation stopping method [5].

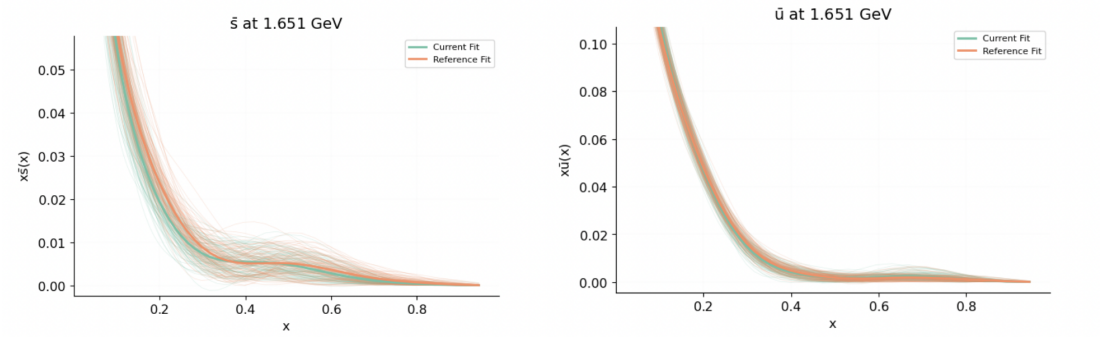
In Figure 2.6 it can be clearly seen how the process of this stopping criterion works. There is in fact a fixed maximum number of possible iterations, and for every iteration it is verified if the positivity of the fitted PDF is fulfilled. Then it is verified if the χ^2_{val} is smaller of the best χ^2 previously computated: if this conditions are fulfilled, then the iteration counter is reset, and the PDF fitted is stored, while if just one of these conditions isn't fulfilled, then the fitted PDF is rejected, and the iteration counter continues.

The importance of using a stopping criterion in the fitting methodology is shown in Figure 2.7, in which we show an example of PDFs fitted with and without using the cross-validation method: we can see how the overfitting leads to a deviation of the PDF fittings from the mean.

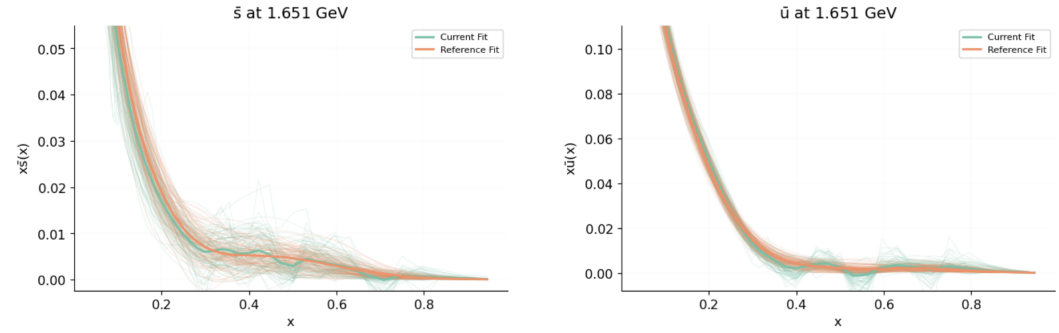
Finally, we must say that there are a few differences between the stopping criterion used in NNPDF4.0 and that of its predecessor used in NNPDF3.1. One of these differences is that the percentage of data that enters the training set has been increased to 75% for all datasets. This is motivated by the observation that the current dataset is so wide that even with just 25% validation overlearning does not occur in practice. Moreover the stopping algorithm in NNPDF4.0 also tracks the positivity requirement so that a fit a priori discarded if the positivity condition is not satisfied. Instead in NNPDF3.1 replicas which were not fulfilling positivity could be generated and had to be discarded a posteriori.

2.3.3 Post-fit selection

After the fitting is done, the mean value of the PDF $f_i(x, Q_0^2)$ is computed with its variance. Then the PDF is given to the APFEL program, which determines the values of the PDF for other values of Q^2, solving the differential equations of perturbative QCD (1.8) and (1.9) presented in Section 1.2. Then we use the post-fit selection, in order to discard PDFs that deviate too much from the average and non integrable PDFs. Finally the PDF is output in the LHAPDF format in order to be used.



(a) $\bar{s}$ and $\bar{u}$ PDF fits with the use of the cross-validation method.



(b) $\bar{s}$ and $\bar{u}$ PDF fits without the use of the cross-validation method.

Figure 2.7: The graphics in Figure 2.7a show an example of PDFs fitted (from two different runcards), using NNPDF4.0 methodology, while the graphics in Figure 2.7a show the same fittings without using a stopping criterion.

Chapter 3

Analysis of the PDF fits

The aim of this thesis is to study in detail the gluon PDF, and determine how the experimental observables, like the production of top and antitop quark pair or jet production, can influence the shape of this distribution. The study of the behavior of the gluon distribution can determine which data cause an impact on its shape.

The analysis of PDF fits is made by comparing the PDFs fitted from different datasets of hadronic cross-section contained in runcards. These runcards are written in YAML lenguage, which is a human-readable data-serialization language. It is commonly used for configuration files and in applications where data is being stored or transmitted.

3.1 First analysis

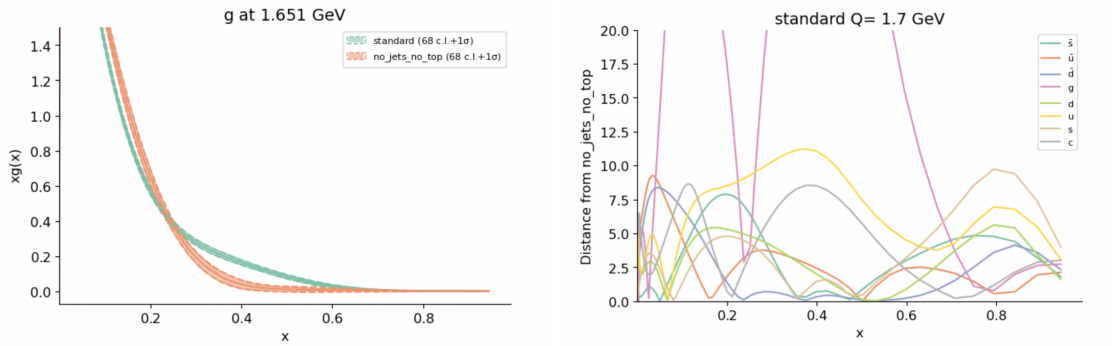
In order to determine which data cause an impact on the gluon PDF, first we take the complete NNPDF4.0 dataset and place it in the runcard standard.yml . After we make a fit of the gluon distribution using this runcard, we remove some datasets from the runcard, linked to some hadronic observables, like the dataset which contain the data of the production of top-antitop quark pair. We then compare the new fit with the old ones, trying to understand how the absence of some datasets influence the gluon distribution.

The first analysis is determined by the first bunch of runcards for which the PDF plots are compared:

- standard.yml : the complete dataset from NNPDF4.0;
- no_jets_no_top.yml : the data from top quark and jet production are excluded;
- no_jets.yml : the jet production data are excluded;
- no_top.yml : the top quark data production are excluded;
- no_jets_no_top_atlasjets.yml : only ATLAS jet production data are included;
- no_jets_no_top_cmsjets.yml : only CMS jet production data are included.

From this analysis we check that indeed the data from top-antitop quark production and jet production have a big influence on the gluon distribution. This fact is given by the comparison between the fits made from the runcards standard.yml and no_jets_no_top.yml, shown in Figure 3.1a. In Figure 3.1b it is presented a comparison of the distance of the two fits, where the distance of two PDFs is defined as follows

$$d(f_1, f_2)(x) \equiv |f_1(x) - f_2(x)| \quad . \tag{3.1}$$



(a) (b)

Figure 3.1: Figure 3.1a shows the plots of the gluon PDFs using the datasets according to the key. Figure 3.1b instead shows the distance of every PDF flavour fitted from no_jets_no_top.yml to the PDFs fitted from standard.yml .

As we can see in Figure 3.1b, the distance of the gluon PDFs fitted from the runcards standard.yml and no_jets_no_top.yml are much greater of the distance of the other fits. It is then understandable that it could be interesting to analyse especially the gluon distribution, in order to understand in what ways the production of top quark and quark jets can influence the interaction between gluons.

For completeness in Figure 3.2 we present also the comparison of the plots of every PDF flavour fitted from the dataset presented earlier.

From the first analysis of the fits, presented in Figure 3.3a and in Figure 3.3b it is possible to make some first considerations. In Figure 3.3a for example, where the gluon PDF fitted from the runcard standard.yml is taken as the reference PDF, we notice that excluding some data, for istance top production data and jet production data, the gluon PDF shows a shifting between the region $0.2 < x < 0.6$: excluding only the top production data, the gluon PDF shifts upwards, while excluding only the jet production data it shifts downwards in that region. Moreover it can be seen that excluding both jet and top production, the gluon PDF shifts downwards in an even more marked way, with respect of the two previous shiftings, as we saw previously. These considerations can be made considering the graphics in Figure 3.4a.

In Figure 3.3b instead, where the gluon PDF made from the runcard no_jets_no_top.yml is taken as the reference PDF, it can be considered that in this case both jet and top production data make the gluon PDF shift up. Furthermore the fact that the two last plots are identical ensures that the jet production data by LHC and CMS are consistent. These considerations can be made considering the graphics in Figure 3.4b.

In conclusion, Figure 3.3a shows that excluding the jets production and top quark production makes the reference PDF shift in two different directions, but on the other hand Figure 3.3b shows that the jet production and top quark production data makes the reference PDF shift in the same direction, that is upwards, but by a different amount, in fact adding the jet production data make the reference PDF shift upwards by a much larger amount.

Then the dataset which would favor a lower gluon distribution in comparison to jet and top pair production, as it is shown in Figure 3.1a, are still unknown. However, an analysis of PDF sets in which datasets are removed one by one in [5] suggests that datasets that have a significant impact on the gluon distribution are the Z boson production ones. We will therefore perform a new analysis also considering these datasets.

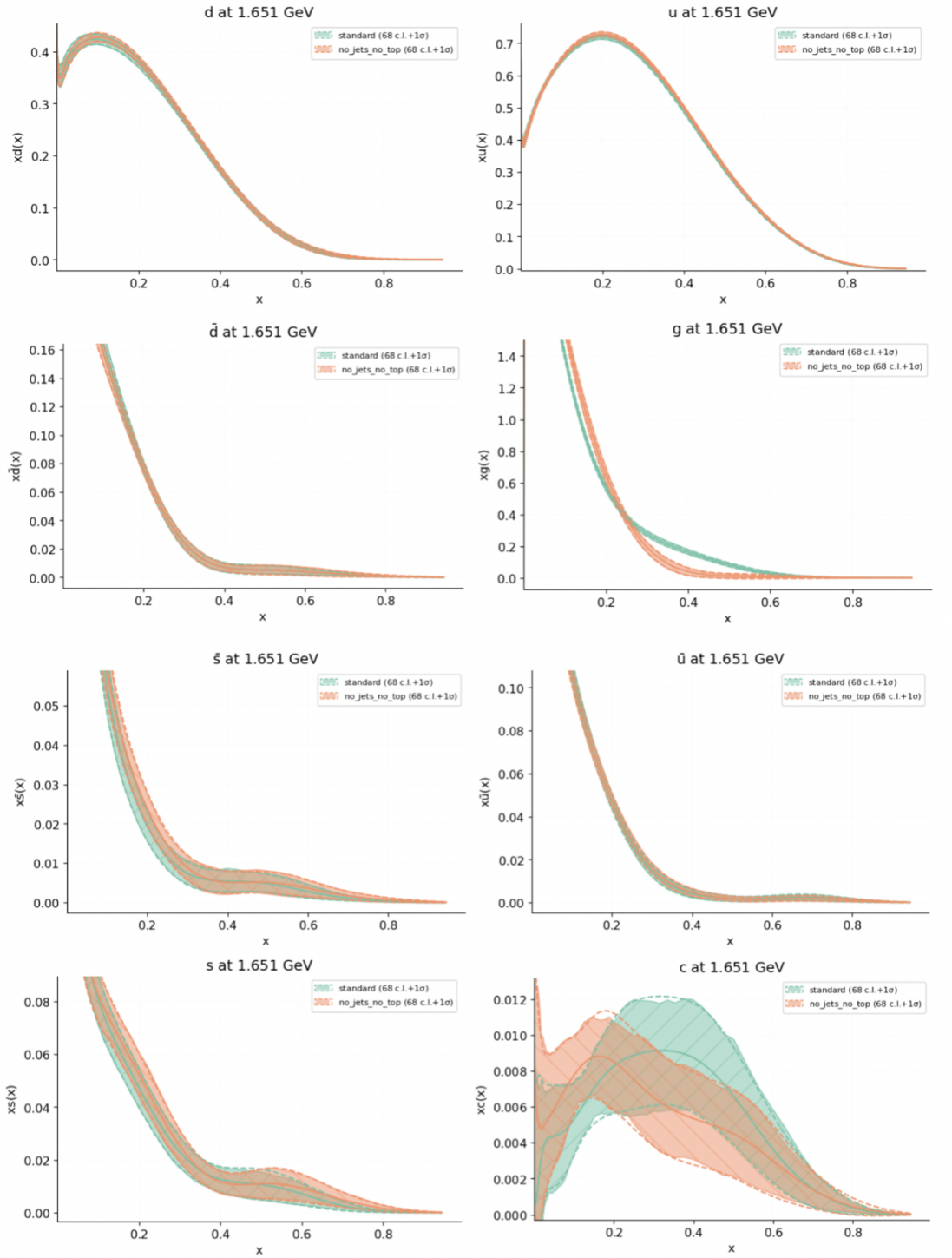
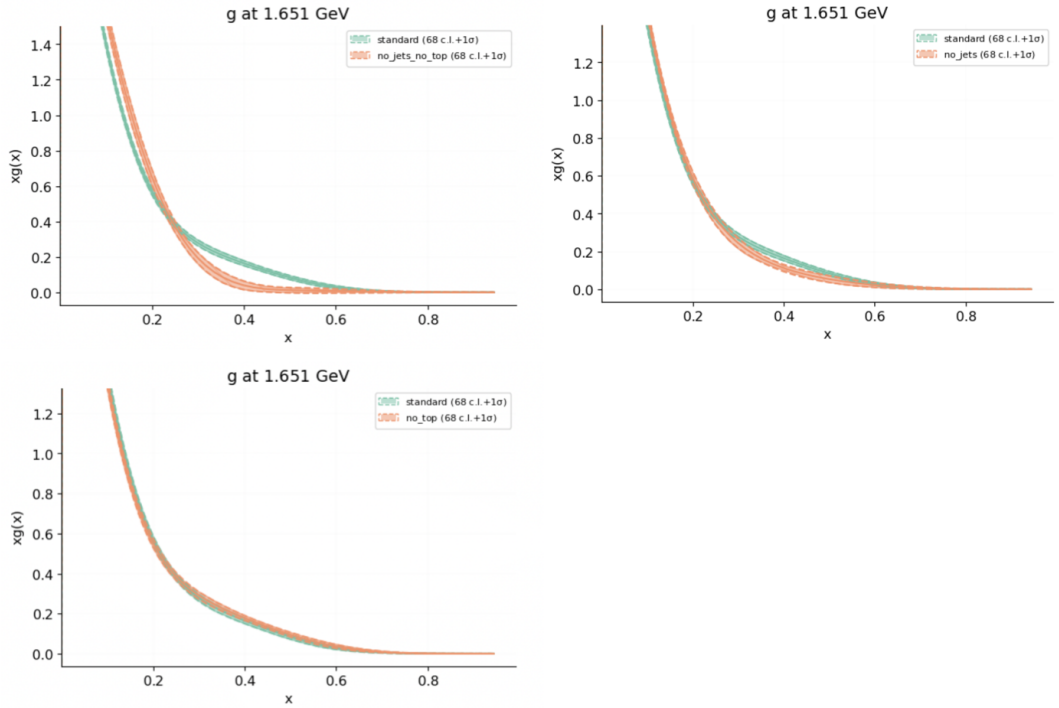
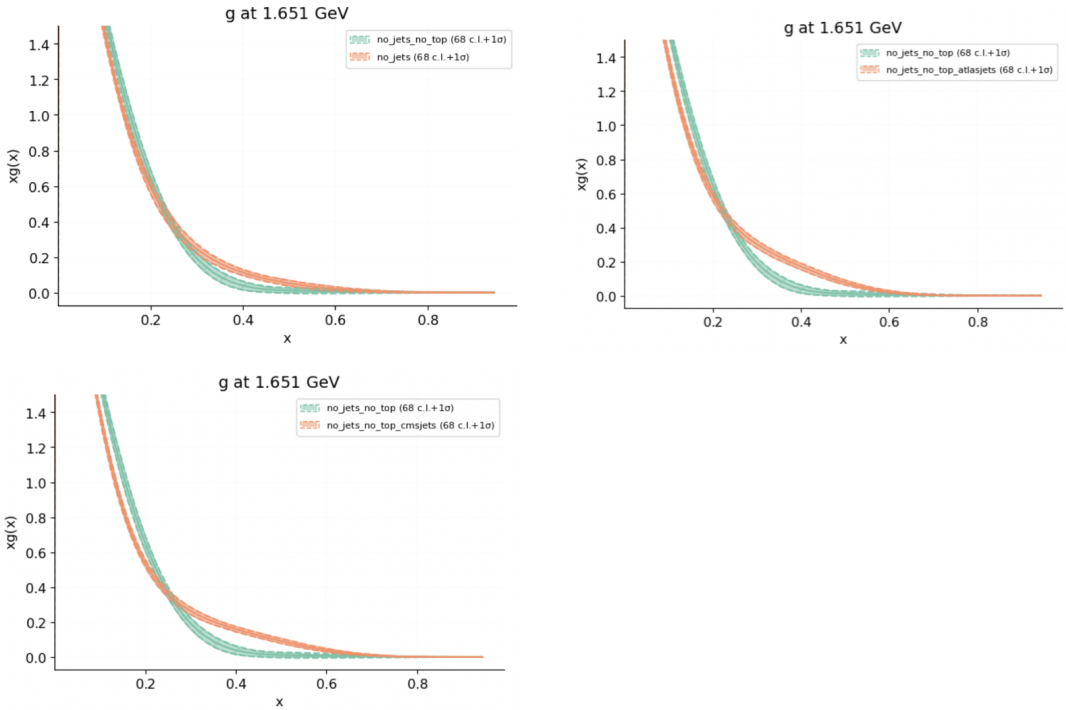


Figure 3.2: Comparison of every flavour PDF plots from the 'first analysis' data.

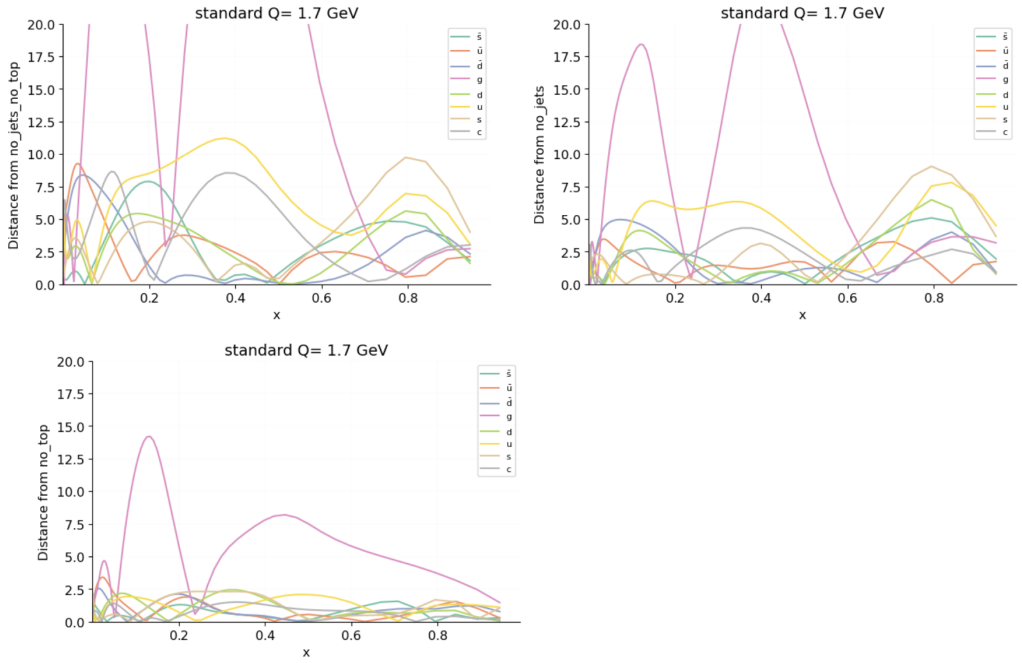


(a) Comparison of gluon PDFs plots in which the reference PDF is the one fitted from the runcard standard.yml .

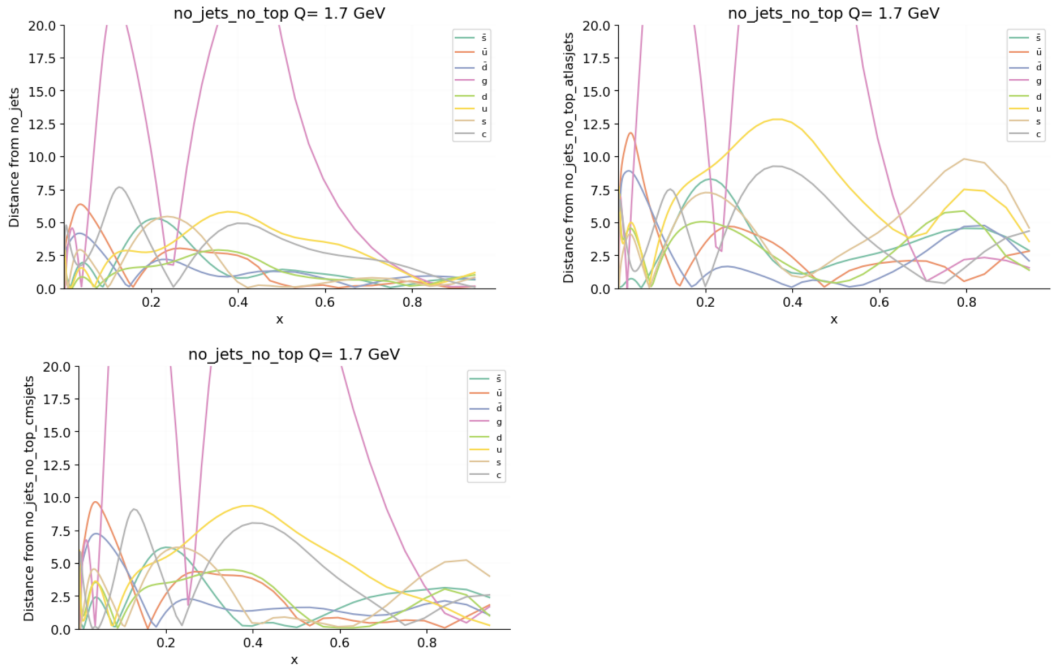


(b) Comparison of gluon PDFs plots in which the reference PDF is the one fitted from the runcard no_jets_no_top.yml .

Figure 3.3: First analysis of the PDF fits comparisons.



(a) Graphics in which we compare the distance of the PDFs fitted from the runcard standard.yml from the PDFs fitted from the runcards no_jets_no_top.yml, no_jets.yml and no_top.yml.



(b) Graphics in which we compare the distance of the PDFs fitted from the runcard no_jets_no_top.yml from the PDFs fitted from the runcards no_jets.yml, no_jets_no_top_atlasjets.yml and no_jets_no_top_cmsjets.yml.

Figure 3.4: Distance between first analysis PDF fits.

3.2 Second analysis

In this second analysis, we again consider the complete NNPDF4.0 dataset, and in addition to the top pair production data and jet production data, we consider also the datasets linked to the Z boson production. Then we again remove these dataset from the runcard standard.yml , and compare the new fits with the old ones.

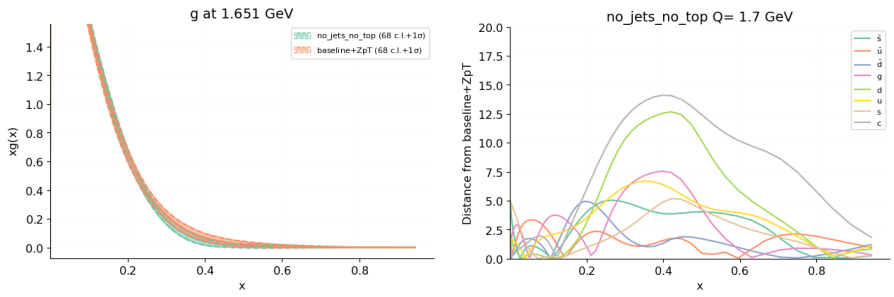
The second bunch of runcards for which the PDF plots are compared, in order to find what hadronic observable can explain the shifting of the PDFs plotted in Figure 3.1a, is the following[1]:

- baseline.yml : the jets, top quark, and Z boson production data from ATLAS and CMS, as well as Z production data from LHCb are excluded;
- baseline+top.yml : top quark production data are re-included in the baseline;
- baseline+jets.yml : jet production data are re-included in the baseline;
- baseline+LHCb.yml : Z boson production data from LHCb are re-included in the baseline;
- baseline+ZpT.yml : Z boson production data (with non-null transversal momentum) from ATLAS and CMS are re-included in the baseline.

In Figure 3.6a we show the comparisons of the PDF fits of these second bunch of runcards, and from these comparisons too it is possible to make some considerations.

First of all it can be seen that the reference gluon PDF fitted in Figure 3.6a, that is the one fitted from the runcard baseline.yml, has a larger variance then the other PDF fits in Figure 3.3 and Figure 3.6a: this is due to the fact that this fit is made removing many dataset from the default runcard baseline.yml.

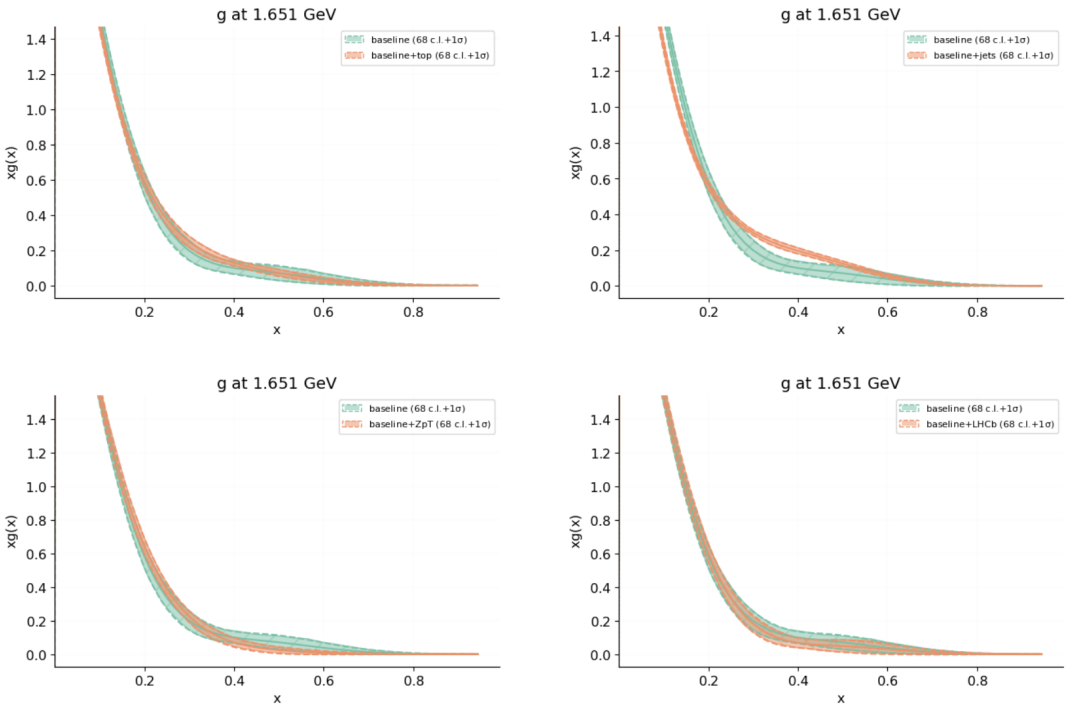
From the analysis of this PDF fits, it can be seen that with the addition of jets production and top quark production data the PDF fits shift upwards in respect to the reference PDF fit from baseline.yml, while with the addition of the Z boson production data the PDF fit shifts downwards. So we can assume that it seems that it is the presence of the Z boson production data with the absence of jets and top quark production data that makes the gluon PDF fit shift downwards and so make this distribution assume such lower values in the region $0.2 < x < 0.6$ compared to the PDF fitted from the runcard standard.yml .



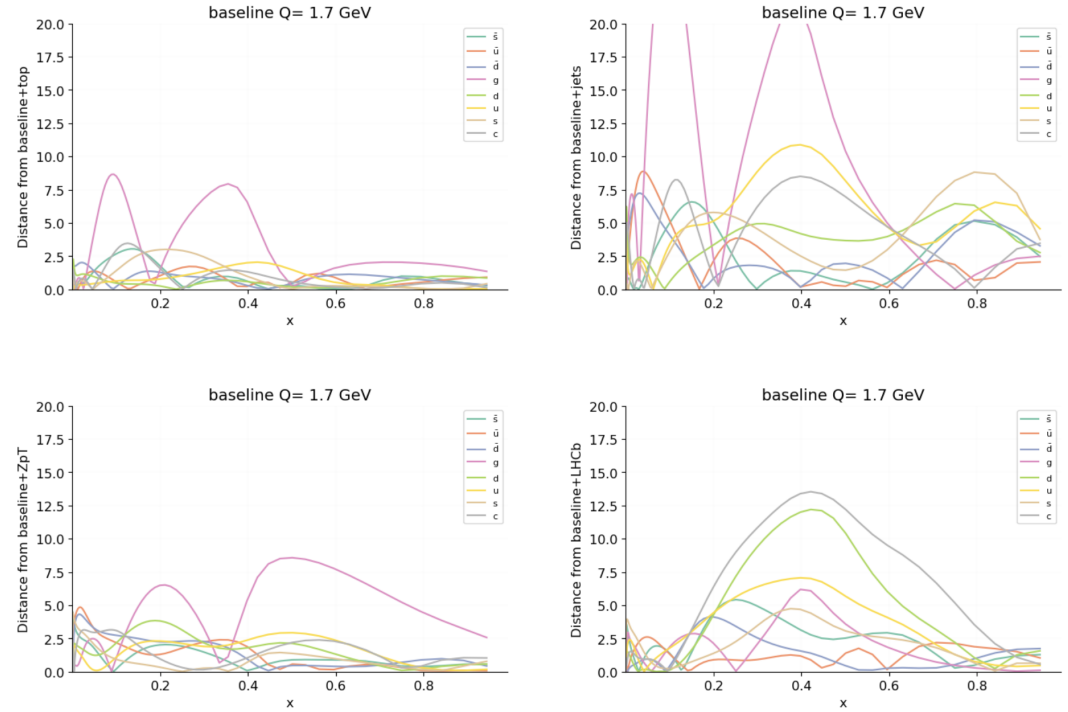
(a) (b)

Figure 3.5: Figure 3.5a shows the comparison between the fits of the gluon PDF using the runcards no_jets_no_top.yml and baselin+ZpT.yml. Figure 3.5b instead shows the distance between the PDF fits.

[1] We consider again the the complete dataset from NNPDF4.0 as reference, that is the runcard standard.yml .



(a) Comparison of gluon PDFs plots in which the reference PDF is the one fitted from the runcard baseline.yml .



(b) Graphics in which we compare the distance of the PDFs fitted from the runcard baseline.yml from the PDFs fitted from the runcards baseline+top.yml, baseline+jets.yml, baseline+ZpT.yml and baseline+LHCb.yml

Figure 3.6: Second analysis of the PDF fits comparisons, and distances between the PDF fits.

Conclusions

In this thesis we presented the problem of determining probability distributions of partons inside the proton, showing how these Parton Distribution Functions encode the structure of strongly-interacting hadrons in the beam of high energy collisions.

We presented then the machine learning method implemented by NNPDF4.0 in order to determine the PDFs strarting from some datasets, which require the gradient descent method and the cross-validation method.

The aim of this thesis was to determine how the experimental observables in high energy scattering experiments can influence the shape of the gluon PDF, and we presented the approach used in finding an answer. First we considered the complete dataset from NNPDF4.0, place it in the runcard standard.yml (the reference runcard), and we made a fit from it. Then we removed some of the datasets from standard.yml linked to the top-antitop pair production and jet production, creating new runcards, and we made a fit from them, comparing the PDFs fitted.

From this first analysis we understood that the the top-antitop quark production data and the jet production data have a huge impact on the gluon distribution, as it seemed that removing both these kind of data from the reference runcard, the PDF fitted from the new runcard no_jets_no_top.yml deviates significantly from the one fitted from standard.yml . However, it seemed also that removing singularly top-antitop production data and jet production data from standard.yml made the gluon PDF shift in different direction from the PDF fitted from the refenrence runcard.

Aware of the fact that the Z boson production data have a significant impact on the gluon distribution, we then performed a new analysis considering also these datasets. So we removed top-antitop production data, jet production data, and Z boson production data from the reference runcard, made a fit from this new runcard (baseline.yml), and then we re-added one by one the different kind of datasets in this runcard, comparing the new fits with the old ones. From this second analysis we found that it is the presence of the Z boson production data with the absence of jets and top quark production data that makes the gluon PDF deviate in a such significant way from the one fitted from the reference runcard.

Of course the deviation of the PDFs fitted from the runcards standard.yml and no_jets_no_top.yml could occur due to a statistical fluctuation, or maybe due to the fact that the jet production data are far more numerous than top-antitop production data, so our conclusion could be wrong. In the future, when we will be equipped of some more data, NNPDF will verify if the deviation of the previously gluon PDFs is due to a lack of data or is due to a truly physically motivation.

Bibliography

[1] Donald H. Perkins. *Introduction to high energy physics*. 2nd ed. Cambridge University press, 2000.

[2] Jian-ping Chen. "Moments of spin structure functions: Sum rules and polarizabilities". In: *International Journal of Modern Physics E* 19 (Jan. 2012). DOI: 10.1142/S0218301310016405.

[3] Guy D. Coughlan, James E. Dodd, and Ben M. Gripaios. *The Ideas of Particle Physics. An Introduction for Scientists*. 2nd ed. Cambridge University press, 2006.

[4] Stefano Forte and Graeme Watt. "Progress in the Determination of the Partonic Structure of the Proton". In: *Ann. Rev. Nucl. Part. Sci.* 63 (2013), pp. 291–328. DOI: 10.1146/annurev-nucl-102212-170607. arXiv: 1301.6754 [hep-ph].

[5] Richard D. Ball et al. "The Path to Proton Structure at One-Percent Accuracy". In: (Sept. 2021). arXiv: 2109.02653 [hep-ph].

[6] Stefano Forte and Stefano Carrazza. "Parton distribution functions". In: (Aug. 2020). arXiv: 2008.12305 [hep-ph].

[7] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. 2016. arXiv: 1603.04467 [cs.DC].

[8] M Aaboud et al. "Precision measurement and interpretation of inclusive W+, W- and Z/gamma* production cross sections with the ATLAS detector". In: *The European Physical Journal C* 77 (June 2017).

[9] M. Aaboud et al. In: *Journal of High Energy Physics* 2018.5 (May 2018). ISSN: 1029-8479. DOI: 10.1007/jhep05(2018)077. URL: http://dx.doi.org/10.1007/JHEP05(2018)077.

[10] M. Aaboud et al. "Measurement of top quark pair differential cross sections in the dilepton channel in pp collisions at $\sqrt{s} = 7, 8$TeV with ATLAS". In: *Physical Review D* 94.9 (Nov. 2016). ISSN: 2470-0029. DOI: 10.1103/physrevd.94.092003. URL: http://dx.doi.org/10.1103/PhysRevD.94.092003.

[11] A. M. Sirunyan et al. "Measurement of double-differential cross sections for top quark pair production in pp collisions at $\sqrt{s} = 8$ TeV and impact on parton distribution functions". In: *The European Physical Journal C* 77.7 (July 2017). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-017-4984-5. URL: http://dx.doi.org/10.1140/epjc/s10052-017-4984-5.

[12] James Currie et al. "Infrared sensitivity of single jet inclusive production at hadron colliders". In: *Journal of High Energy Physics* 2018.10 (Oct. 2018). ISSN: 1029-8479. DOI: 10.1007/jhep10(2018)155. URL: http://dx.doi.org/10.1007/JHEP10(2018)155.

[13] G. Aad et al. "Measurement of the inclusive isolated prompt photon cross section in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector". In: *Journal of High Energy Physics* 2016.8 (Aug. 2016). ISSN: 1029-8479. DOI: 10.1007/jhep08(2016)005. URL: http://dx.doi.org/10.1007/JHEP08(2016)005.

[14] M. Aaboud et al. "Measurement of the cross section for inclusive isolated-photon production in *pp* collisions at $\sqrt{s} = 13\,\text{TeV}$ using the ATLAS detector". In: *Physics Letters B* 770 (July 2017), pp. 473–493. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2017.04.072. URL: http://dx.doi.org/10.1016/j.physletb.2017.04.072.

[15] G. Aad et al. "Comprehensive measurements oft-channel single top-quark production cross sections at $\sqrt{s} = 7\,\text{TeV}$ with the ATLAS detector". In: *Physical Review D* 90.11 (Dec. 2014). ISSN: 1550-2368. DOI: 10.1103/physrevd.90.112006. URL: http://dx.doi.org/10.1103/PhysRevD.90.112006.

[16] M. Aaboud et al. "Fiducial, total and differential cross-section measurements of t-channel single top-quark production in pp collisions at $\sqrt{s} = 8\,\text{TeV}$ using data collected by the ATLAS detector". In: *The European Physical Journal C* 77.8 (Aug. 2017). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-017-5061-9. URL: http://dx.doi.org/10.1140/epjc/s10052-017-5061-9.

[17] M. Aaboud et al. "Measurement of the inclusive cross-sections of single top-quark and top-antiquark t-channel production in pp collisions at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS detector". In: *Journal of High Energy Physics* 2017.4 (Apr. 2017). ISSN: 1029-8479. DOI: 10.1007/jhep04(2017)086. URL: http://dx.doi.org/10.1007/JHEP04(2017)086.

[18] S. Chatrchyan et al. "Measurement of the single-top-quark t-channel cross section in pp collisions at $\sqrt{s} = 7\,\text{TeV}$". In: *Journal of High Energy Physics* 2012.12 (Dec. 2012). ISSN: 1029-8479. DOI: 10.1007/jhep12(2012)035. URL: http://dx.doi.org/10.1007/JHEP12(2012)035.

[19] V. Khachatryan et al. "Measurement of the t-channel single-top-quark production cross section and of the |Vtb| CKM matrix element in pp collisions at $\sqrt{s} = 8\,\text{TeV}$". In: *Journal of High Energy Physics* 2014.6 (June 2014). ISSN: 1029-8479. DOI: 10.1007/jhep06(2014)090. URL: http://dx.doi.org/10.1007/JHEP06(2014)090.

[20] A.M. Sirunyan et al. "Cross section measurement of t-channel single top quark production in pp collisions at $\sqrt{s} = 13\,\text{TeV}$". In: *Physics Letters B* 772 (Sept. 2017), pp. 752–776. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2017.07.047. URL: http://dx.doi.org/10.1016/j.physletb.2017.07.047.

[21] Alessandro Candido, Stefano Forte, and Felix Hekhorn. "Can $\overline{\text{MS}}$ parton distributions be negative?" In: *Journal of High Energy Physics* 2020.11 (Nov. 2020). ISSN: 1029-8479. DOI: 10.1007/jhep11(2020)129. URL: http://dx.doi.org/10.1007/JHEP11(2020)129.

[22] Richard D. Ball et al. "Parton distributions for the LHC run II". In: *Journal of High Energy Physics* 2015.4 (Apr. 2015). ISSN: 1029-8479. DOI: 10.1007/jhep04(2015)040. URL: http://dx.doi.org/10.1007/JHEP04(2015)040.

[23] R.G. Roberts. *The structure of the proton: Deep inelastic scattering*. Cambridge University press, 1990.

[24] James Bergstra, Daniel Yamins, and David Cox. "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures". In: Proceedings of Machine Learning Research 28.1 (2013). Ed. by Sanjoy Dasgupta and David McAllester. URL: https://proceedings.mlr.press/v28/bergstra13.html.